ArcFace:
Additive Angular Margin Loss for Deep Face Recognition

Jiankang Deng, Jia Guo, Niannan Xue, Stefanos Zafeiriou

Imperial College London, Insight Face

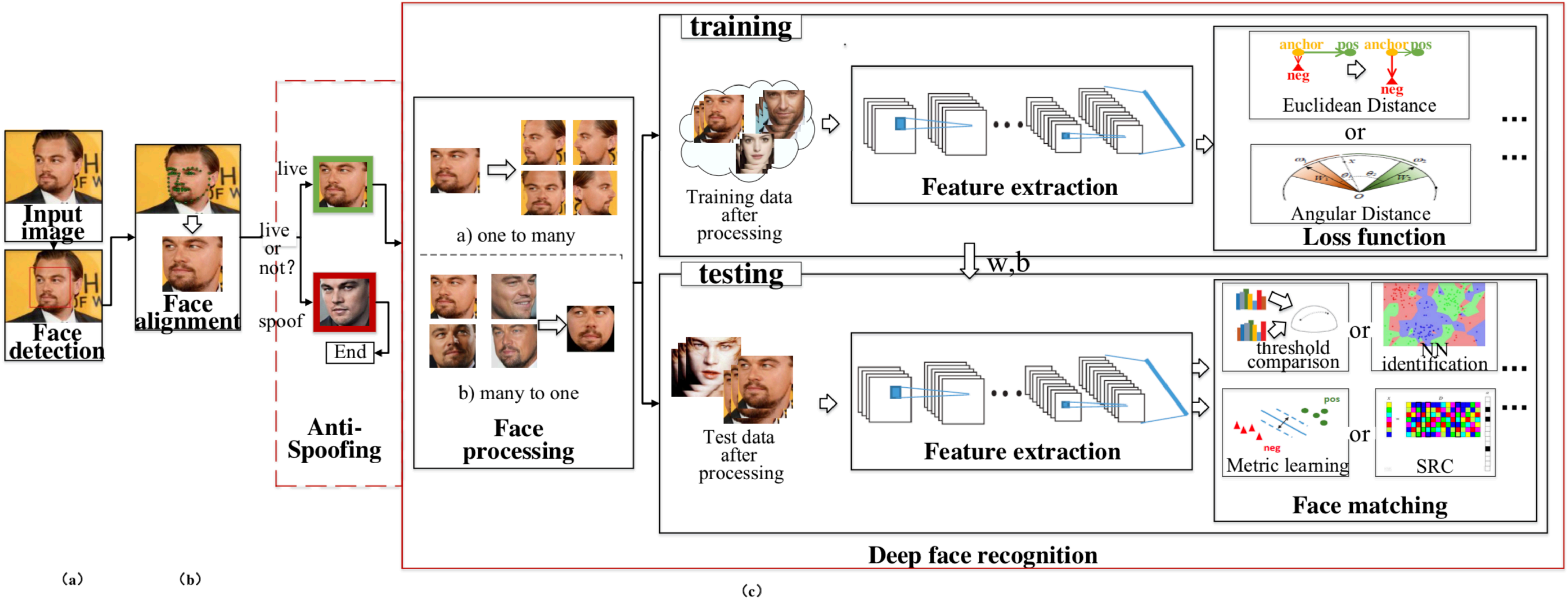
arXiv:1801.07698

Sungman, Cho.

Introduction

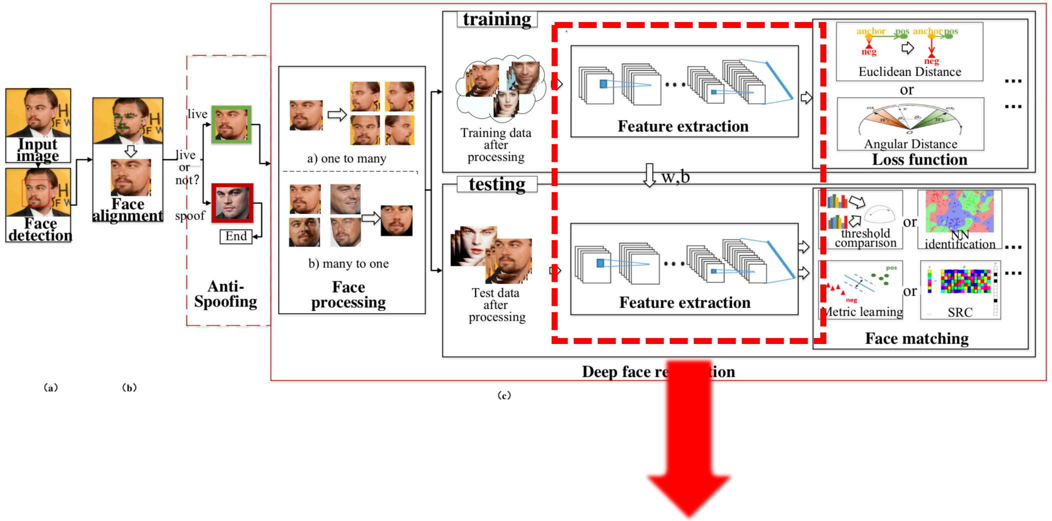
Dive into FR(Face Recognition)

Deep FR(Face Recognition) System



Ref. M.Wang, "Deep Face Recognition", arXiv:1804.06655

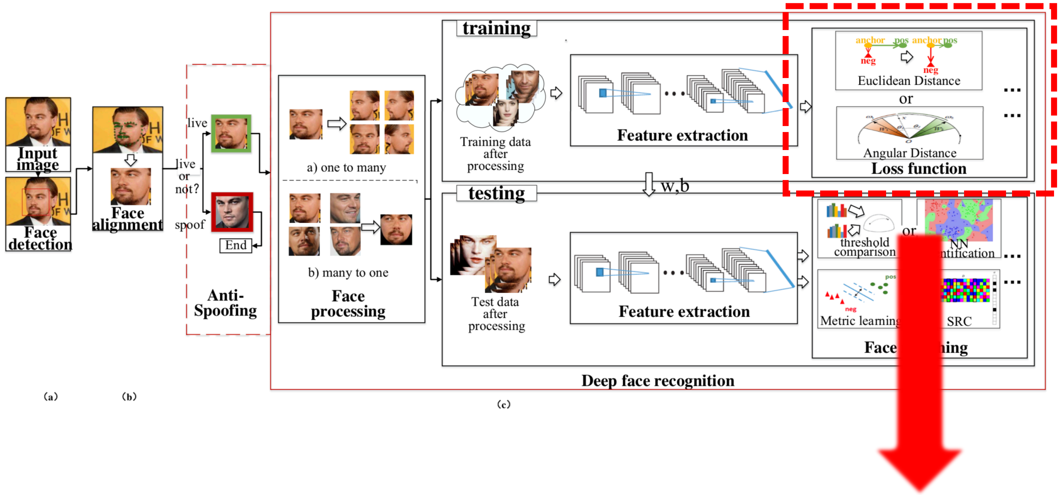
Feature Extraction Network



Network Architectures	Subsettings
backbone network	mainstream architectures: AlexNet [140], [139], [144], VGGNet [123], [116], [224], GoogleNet [204], [144], ResNet [106], [224], SENet [20]
	special architectures [187], [188], [157], [34], [194]
	joint alignment-representation architectures [64], [186], [237], [29]
multiple networks	multipose [87], [115], [211], [175], multipatch [105], [239], [46], [155], [156], [152] [185], multitask [131]

Ref. M.Wang, "Deep Face Recognition", arXiv:1804.06655

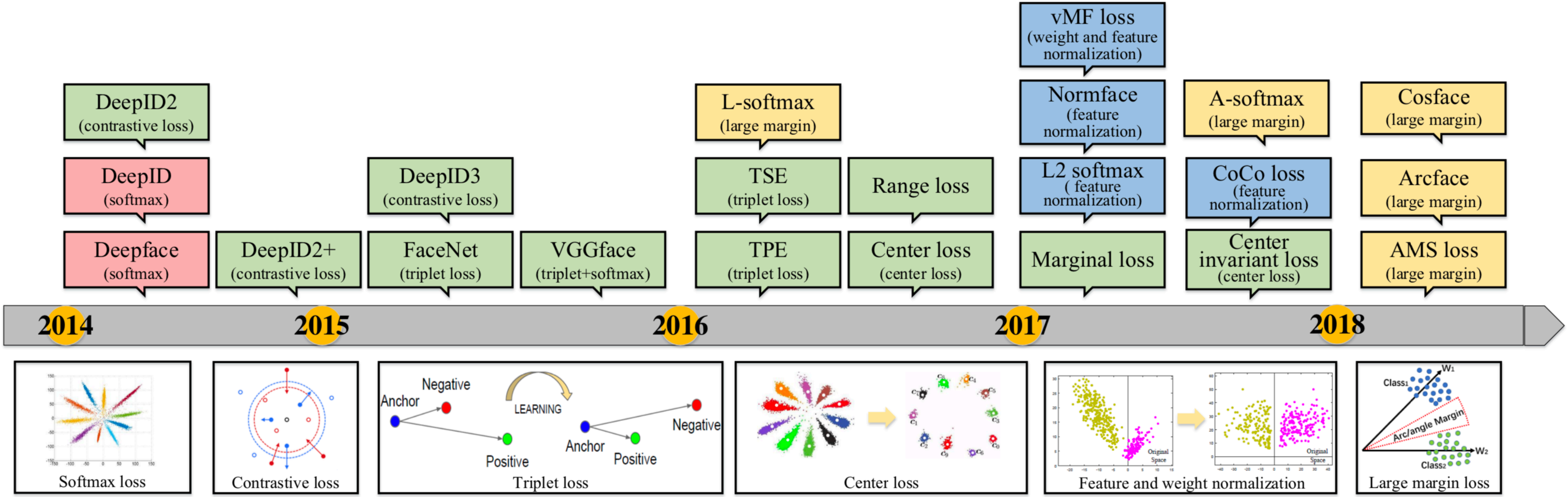
Loss Function



Loss Functions	Brief Description
Euclidean-distance-based loss	compressing intra-variance and enlarging inter-variance based on Euclidean distance. [152], [185], [153], [181], [191], [224], [144], [123], [140], [139], [105], [28]
angular/cosine-margin-based loss	making learned features potentially separable with larger angular/cosine distance. [107], [106], [170], [38], [172], [108]
softmax loss and its variations	modifying the softmax loss to improve performance. [129], [171], [61], [111] [128], [23], [62]

Ref. M.Wang, “Deep Face Recognition”, arXiv:1804.06655

Loss Function



Ref. M.Wang, “Deep Face Recognition”, arXiv:1804.06655

Main stream of FR

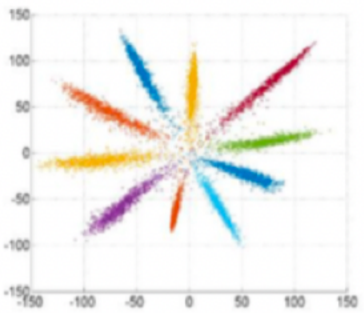
- **Softmax**

  Train a multi-class classifier which can separate different identities in the training set

- **Triplet loss**

  Learn directly an embedding

Main stream of FR



Softmax loss

- **Softmax**

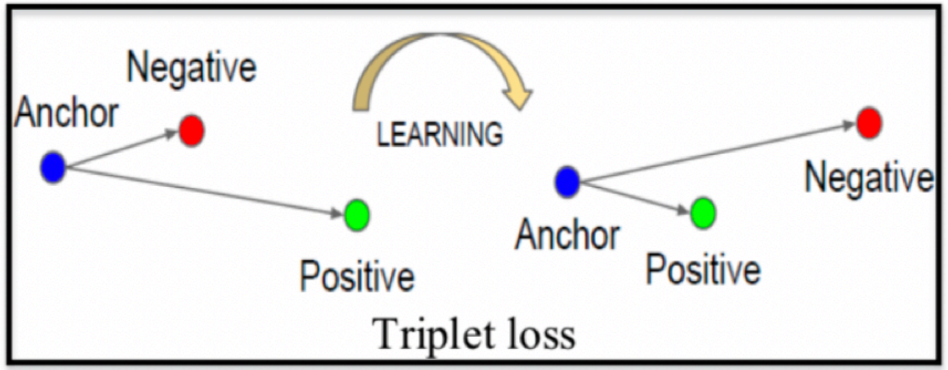
Train a multi-class classifier which can separate different identities in the training set

[cons]

- The size of the linear transformation matrix $W \in \mathbb{R}^{d \times n}$ increases linearly.
- The learned features are separable for the closed-set classification problem.
  (not discriminative enough for the opn-set face recognition problem.)

Main stream of FR

- **Triplet loss**



Learn directly an embedding

[cons]

- Combinatorial explosion in the number of face triplets.
- Semi-hard sample mining is a quite difficult problem for effective training.

Variants

- **Center loss (ECCV, 2016)**
- **Sphereface (CVPR, 2017)**
- **ArcFace (arXiv:1801.07698)**

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std(%)
DeepFace [160]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35±0.25
DeepID2 [152]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15±0.13
DeepID3 [153]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53±0.10
FaceNet [144]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63±0.09
Baidu [105]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [123]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [188]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [181]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [107]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [224]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [129]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [171]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [111]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [62]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [39]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [106]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [128]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [170]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [172]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [38]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [235]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

Ref. M.Wang, “Deep Face Recognition”, arXiv:1804.06655

Variants

- **Center loss (ECCV, 2016)**
- Sphereface (CVPR, 2017)
- ArcFace (arXiv:1801.07698)

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std(%)
DeepFace [160]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35±0.25
DeepID2 [152]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15±0.13
DeepID3 [153]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53±0.10
FaceNet [144]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63±0.09
Baidu [105]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [123]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [188]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [181]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [107]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [224]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [129]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [171]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [111]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [62]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [39]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [106]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [128]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [170]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [172]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [38]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [235]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

Ref. M.Wang, "Deep Face Recognition", arXiv:1804.06655

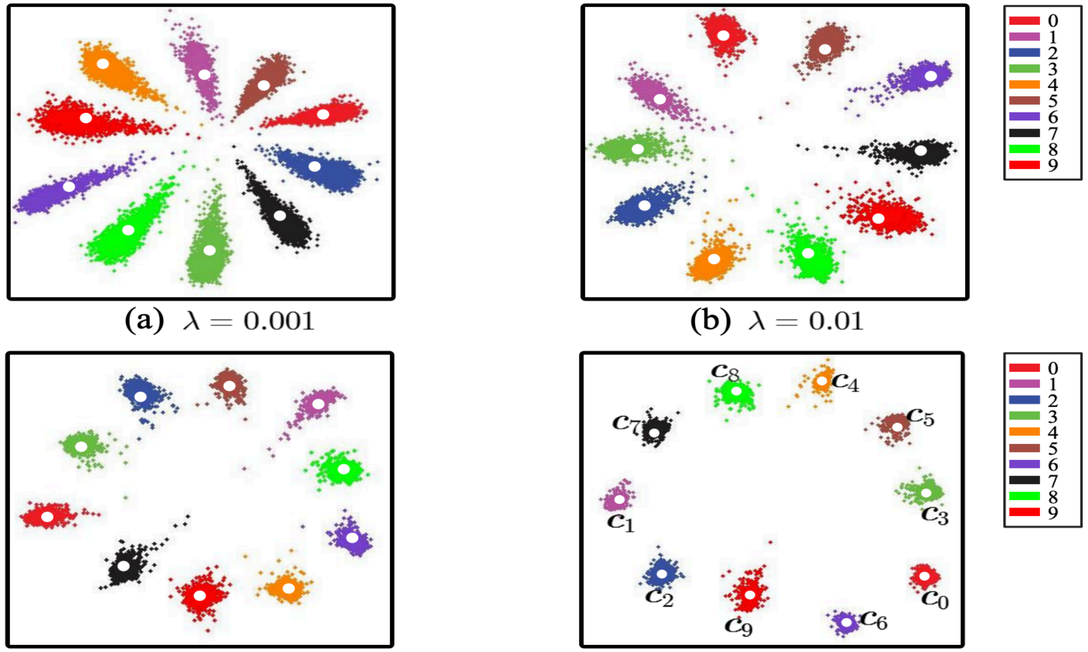
Center Loss

- The Euclidean distance between each feature vector and its class center

- To obtain intra-class compactness & inter-class dispersion

- Updating the actual centers during training is difficult.

  (the number of face classes has dramatically increased)



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$ (c) $\lambda = 0.1$ (d) $\lambda = 1$

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C$$

$$= -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

Ref. Y.Wen, “A Discriminative Feature Learning Approach for Deep Face Recognition”, ECCV, 2016.

Variants

- Center loss (ECCV, 2016)
- **Sphereface (CVPR, 2017)**
- ArcFace (arXiv:1801.07698)

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std(%)
DeepFace [160]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35±0.25
DeepID2 [152]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15±0.13
DeepID3 [153]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53±0.10
FaceNet [144]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63±0.09
Baidu [105]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [123]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [188]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [181]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [107]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [224]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [129]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [171]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [111]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [62]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [39]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [106]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [128]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [170]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [172]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [38]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [235]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

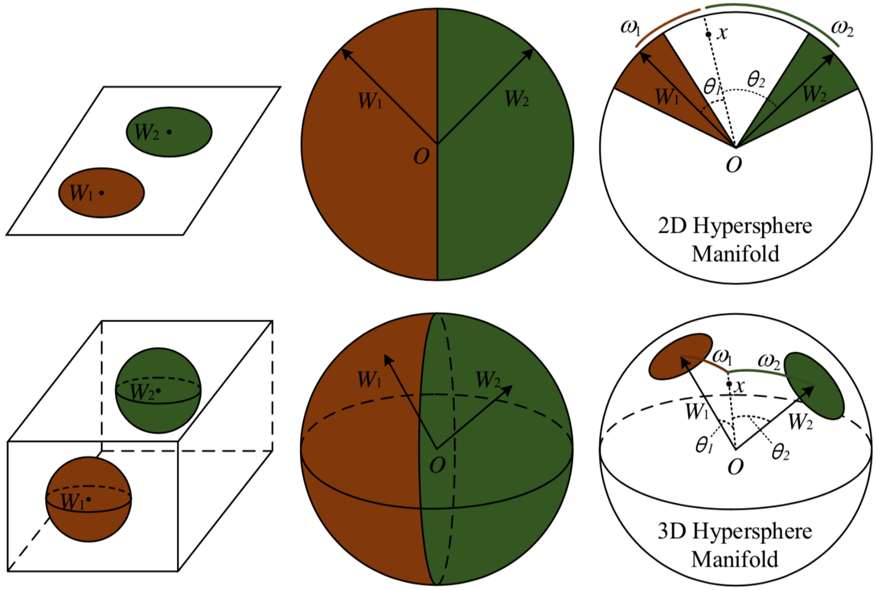
Ref. M.Wang, "Deep Face Recognition", arXiv:1804.06655

Sphere Face

- Introduce angular margin

- Loss-function required a series of approximations in order to be computed.

- Approximation resulted in an unstable training of the network



Euclidean Margin Loss　Modified Softmax Loss　A-Softmax Loss (m≥2)

Loss Function	Decision Boundary
Softmax Loss	$(\boldsymbol{W}_1 - \boldsymbol{W}_2)\boldsymbol{x} + b_1 - b_2 = 0$
Modified Softmax Loss	$\|\boldsymbol{x}\|(\cos\theta_1 - \cos\theta_2) = 0$
A-Softmax Loss	$\|\boldsymbol{x}\|(\cos m\theta_1 - \cos\theta_2) = 0$ for class 1 $\|\boldsymbol{x}\|(\cos\theta_1 - \cos m\theta_2) = 0$ for class 2

Ref. W.Liu, "SphereFace: Deep Hypersphere Embedding for Face Recognition", CVPR 2017

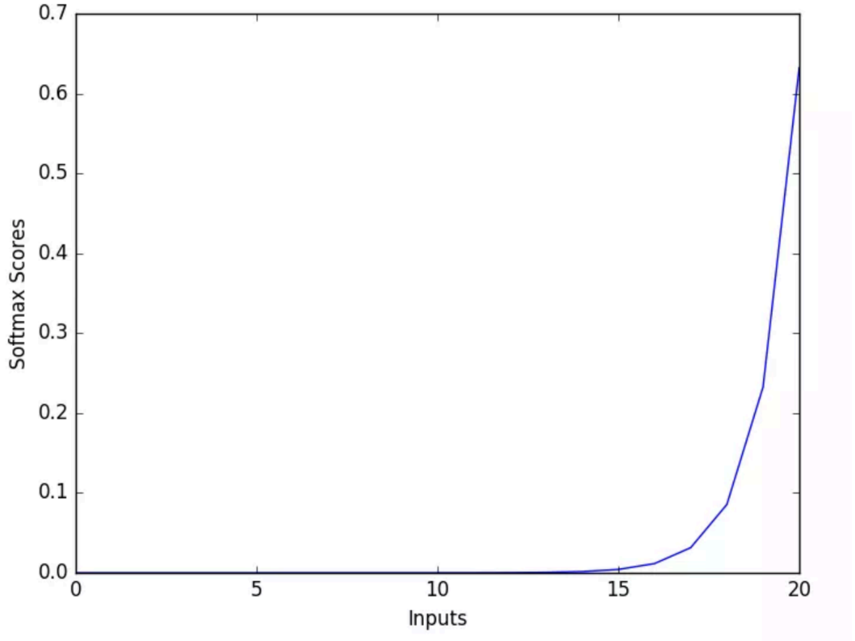
Methodology

1. ArcFace

Revisiting the Softmax

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}},$$

$x_i \in \mathbb{R}^d$: deep feature of the i -th sample, belonging to the y_i-th class.
$W_j \in \mathbb{R}^d$: j -th column of the weight $W \in \mathbb{R}^{d \times n}$
$b_j \in \mathbb{R}^n$: bias term

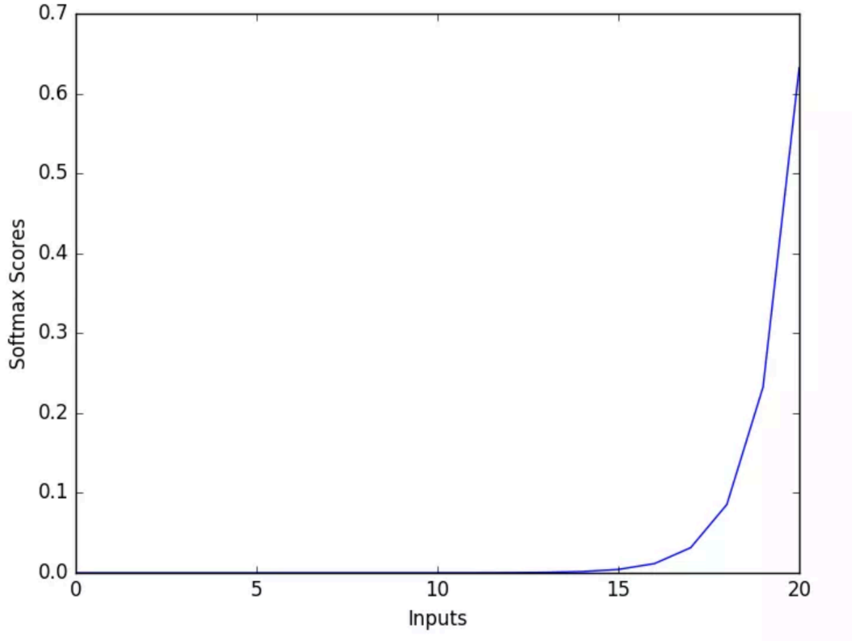


Softmax Graph

Revisiting the Softmax

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}},$$

$x_i \in \mathbb{R}^d$: deep feature of the i -th sample, belonging to the y_i-th class.
$W_j \in \mathbb{R}^d$: j -th column of the weight $W \in \mathbb{R}^{d \times n}$
$b_j \in \mathbb{R}^n$: bias term



Softmax Graph

The softmax loss function does not explicitly optimize the feature embedding to enforce higher similarity for intra-class samples and diversity for inter-class samples.

Feature Embedding

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}},$$

$(b_j = 0)$ $\quad W_j^T x_i = \left\|W_j^T\right\| \|x_i\| \cos\theta_j$ $\quad \theta_j$: angle between W_j and the x_i

$(l_2\ normalization)$ $\quad \underbrace{\|x_i\| = s,}_{+\ \text{rescale}}$ $\quad \left\|W_j\right\| = 1$

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}.$$

Feature Embedding

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}},$$

$(b_j = 0)$ $\quad W_j^T x_i = \|W_j^T\| \|x_i\| \cos\theta_j$

$(l_2\ normalization)$ $\quad \underline{\|x_i\| = s,}\ \ \|W_j\| = 1$

+ rescale

θ_j : angle between W_j and the x_i

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}.$$

Feature Embedding

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}},$$

Make the predictions *only depend on the angle between the feature and the weight*

$(b_j = 0)$ $\quad W_j^T x_i = \|W_j^T\|\|x_i\|\cos\theta_j$

$(l_2\ normalization)$ $\quad \underbrace{\|x_i\| = s,}_{+\ \text{rescale}} \quad \|W_j\| = 1$

θ_j : angle between W_j and the x_i

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}.$$

Feature Embedding

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}},$$

Make the predictions only depend on the angle between the feature and the weight

$(b_j = 0)$ $\quad W_j^T x_i = \|W_j^T\|\|x_i\|\cos\theta_j$

$(l_2\ normalization)$ $\quad \underline{\|x_i\| = s,}\quad \|W_j\| = 1$

+ rescale

θ_j : angle between W_j and the x_i

The learned embedding features are thus *distributed on a hypersphere with a radius of s*

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}.$$

Feature Embedding

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}.$$

Add an additive angular margin penalty m between x_i and W_{y_i}

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}.$$

Feature Embedding

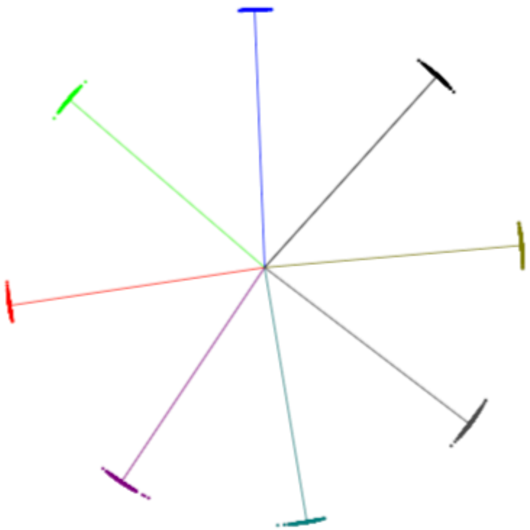
$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}},$$

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}.$$

(a) Softmax

Line: center direction

Dots: samples



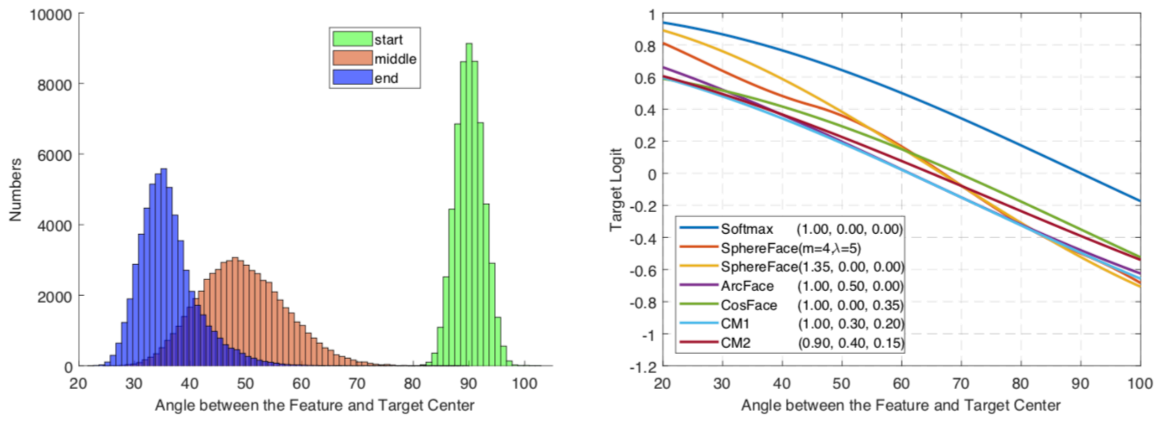
(b) ArcFace

Numerical Similarity.

- **SphereFace, ArcFace, CosFace**

Three different kinds of margin penalty

- Multiplicative angular margin m_1
- Additive angular margin m_2
- Additive cosine margin m_3

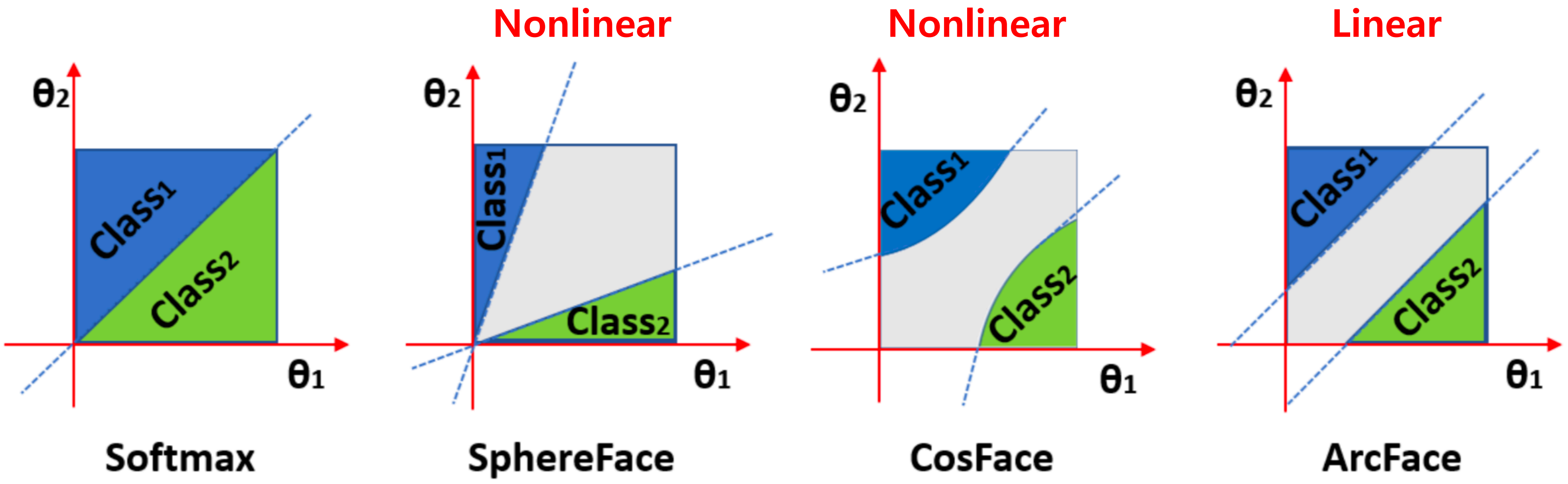


(a) θ_j Distributions (b) Target Logits Curves

From the view of numerical analysis,
All enforce the intra-class compactness and inter-class diversity by penalizing the target logit

Geometric Difference

- SphereFace, ArcFace, CosFace



The minor difference in margin designs can have “butterfly effect” on the model training.

Compare with Other Losses

- **Intra-Loss**

  Improve the intra compactness by decreasing the angle/arc between the sample and the ground truth center

$$L_5 = L_2 + \frac{1}{\pi N} \sum_{i=1}^{N} \theta_{y_i}.$$

- **Inter-Loss**

  enhance inter-class discrepancy by increasing the angle/arc between different centers.

$$L_6 = L_2 - \frac{1}{\pi N (n-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq y_i}^{n} \arccos(W_{y_i}^T W_j). \quad (6)$$

- **Triplet-Loss**

  enlarge the angle/arc margin between triplet samples.

$$\arccos(x_i^{pos} x_i) + m \leq \arccos(x_i^{neg} x_i).$$

Experiments

1. Datasets, Implementation Details.

Datasets & Implementation Details

Datasets	#Identity	#Image/Video
CASIA [43]	10K	0.5M
VGGFace2 [6]	9.1K	3.3M
MS1MV2	85K	5.8M
MS1M-DeepGlint [2]	87K	3.9M
Asian-DeepGlint [2]	94 K	2.83M
LFW [13]	5,749	13,233
CFP-FP [30]	500	7,000
AgeDB-30 [22]	568	16,488
CPLFW [48]	5,749	11,652
CALFW [49]	5,749	12,174
YTF [40]	1,595	3,425
MegaFace [15]	530 (P)	1M (G)
IJB-B [39]	1,845	76.8K
IJB-C [21]	3,531	148.8K
Trillion-Pairs [2]	5,749 (P)	1.58M (G)
iQIYI-VID [20]	4,934	172,835

- **By utilizing five facial points, generate the normalized face crop (112x112)**
- **Employ ResNet50, ResNet100**
- **After the last Conv.layer, explore [BN-Dropout-FC-BN] structure to get the final 512-D embeddings**
- **Set feature scale s=64, angular margin m = 0.5**
- **batch size = 512, lr=0.1(20K: 0.01, 28K: 0.001), momentum=0.9**
- **weight decay = 5e-4**
- **Finished at 32K iteration, Using MxNet, NVIDIA Tesla P40(x4),**

Verification Results

Loss Functions	LFW	CFP-FP	AgeDB-30
ArcFace (0.4)	99.53	95.41	94.98
ArcFace (0.45)	99.46	95.47	94.93
ArcFace (0.5)	**99.53**	**95.56**	**95.15**
ArcFace (0.55)	99.41	95.32	95.05
SphereFace [18]	99.42	-	-
SphereFace (1.35)	99.11	94.38	91.70
CosFace [37]	99.33	-	-
CosFace (0.35)	99.51	95.44	94.56
CM1 (1, 0.3, 0.2)	99.48	95.12	94.38
CM2 (0.9, 0.4, 0.15)	99.50	95.24	94.86
Softmax	99.08	94.39	92.33
Norm-Softmax (NS)	98.56	89.79	88.72
NS+Intra	98.75	93.81	90.92
NS+Inter	98.68	90.67	89.50
NS+Intra+Inter	98.73	94.00	91.41
Triplet (0.35)	98.98	91.90	89.98
ArcFace+Intra	99.45	95.37	94.73
ArcFace+Inter	99.43	95.25	94.55
ArcFace+Intra+Inter	99.43	95.42	95.10
ArcFace+Triplet	99.50	95.51	94.40

Experiments

2. Angle Statistics

The angle statistics (Preview)

	NS	ArcFace	IntraL	InterL	TripletL
W-EC	44.26	14.29	8.83	46.85	-
W-Inter	69.66	71.61	31.34	75.66	-
Intra1	50.50	38.45	17.50	52.74	41.19
Inter1	59.23	65.83	24.07	62.40	50.23
Intra2	33.97	28.05	12.94	35.38	27.42
Inter2	65.60	66.55	26.28	67.90	55.94

W-EC : mean of angles between W_j and the corresponding embedding feature center
W-Inter : mean of minimum angels between W_j's.
Intra1, Intra2 : mean of angles between x_i and the embedding feature center on CASIA, LFW
Inter1, Inter2 : mean of minimum angles between embedding feature center on CASIA, LFW

The angle statistics (Norm-Softmax)

W-EC : mean of angles between W_j and the corresponding embedding feature center

	NS	ArcFace	IntraL	InterL	TripletL
W-EC	44.26	14.29	8.83	46.85	-
W-Inter	69.66	71.61	31.34	75.66	-
Intra1	50.50	38.45	17.50	52.74	41.19
Inter1	59.23	65.83	24.07	62.40	50.23
Intra2	33.97	28.05	12.94	35.38	27.42
Inter2	65.60	66.55	26.28	67.90	55.94

Obvious deviation (44.26) between W_j and the embedding feature center

W_j can't absolutely represent the inter-class discrepancy on training data.

The angle statistics (Intra-Loss)

GOAL : Compress Intra-class & Increase Inter-class discrepancy !!!

	NS	ArcFace	IntraL	InterL	TripletL
W-EC	44.26	14.29	8.83	46.85	-
W-Inter	69.66	71.61	31.34	75.66	-
Intra1	50.50	38.45	17.50	52.74	41.19
Inter1	59.23	65.83	24.07	62.40	50.23
Intra2	33.97	28.05	12.94	35.38	27.42
Inter2	65.60	66.55	26.28	67.90	55.94

Intra-Loss can effectively compress intra-class variations but also brings in smaller inter-class angles.

The angle statistics (Inter-Loss)

GOAL : Compress Intra-class & Increase Inter-class discrepancy !!!

	NS	ArcFace	IntraL	InterL	TripletL
W-EC	44.26	14.29	8.83	46.85	-
W-Inter	69.66	71.61	31.34	75.66	-
Intra1	50.50	38.45	17.50	52.74	41.19
Inter1	59.23	65.83	24.07	62.40	50.23
Intra2	33.97	28.05	12.94	35.38	27.42
Inter2	65.60	66.55	26.28	67.90	55.94

Inter-Loss can slightly increase inter-class discrepancy on both W and embedding network, but also raise intra-class angles.

The angle statistics (Inter-Loss)

GOAL : Compress Intra-class & Increase Inter-class discrepancy !!!

	NS	ArcFace	IntraL	InterL	TripletL
W-EC	44.26	14.29	8.83	46.85	-
W-Inter	69.66	71.61	31.34	75.66	-
Intra1	50.50	38.45	17.50	52.74	41.19
Inter1	59.23	65.83	24.07	62.40	50.23
Intra2	33.97	28.05	12.94	35.38	27.42
Inter2	65.60	66.55	26.28	67.90	55.94

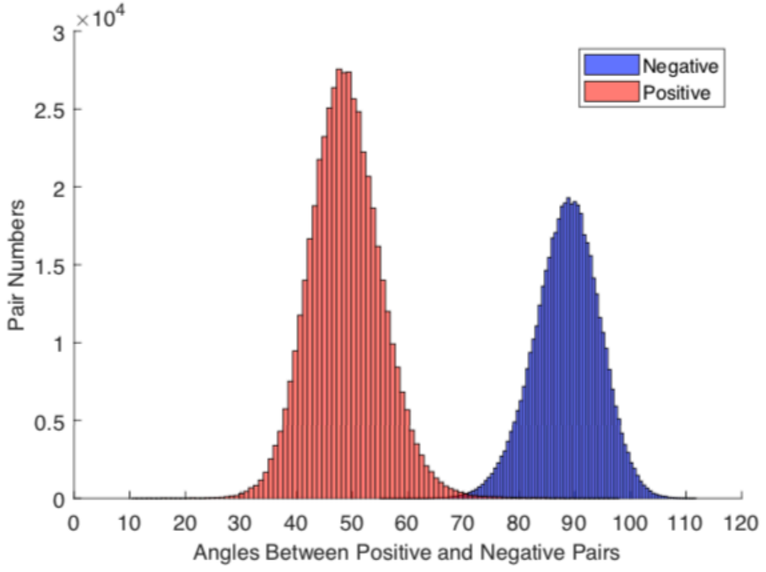
Triplet-Loss has similar intra-class compactness
but inferior inter-class discrepancy compared to ArcFace.

Experiments

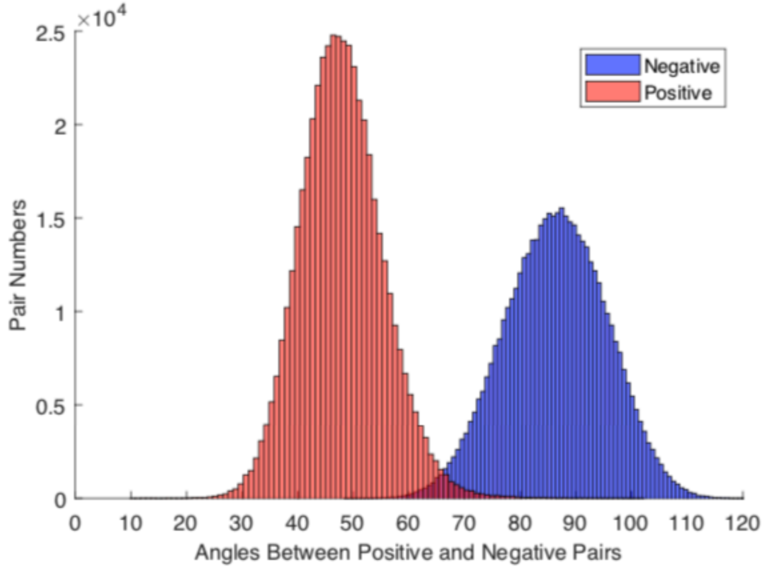
3. Tests

ArcFace vs Triplet-Loss

LFW datasets



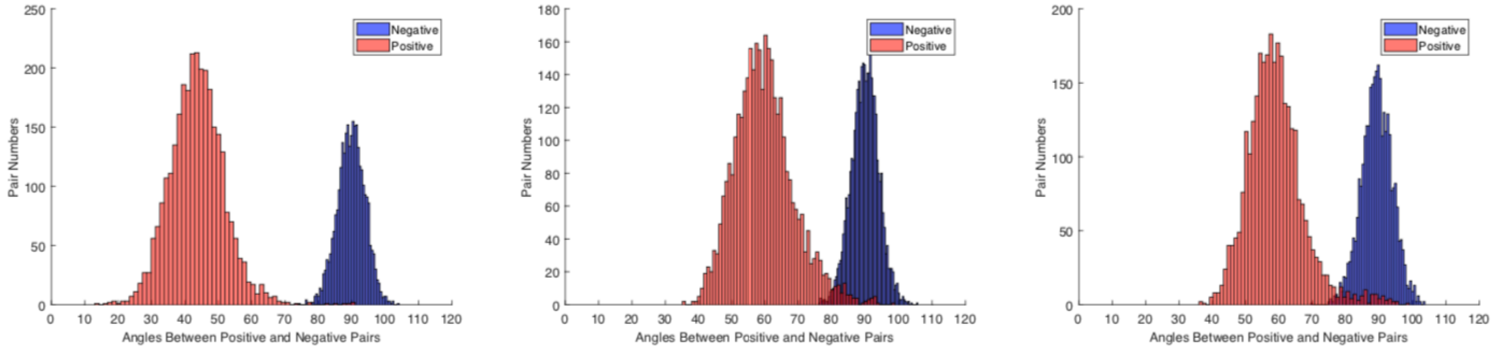
(a) ArcFace



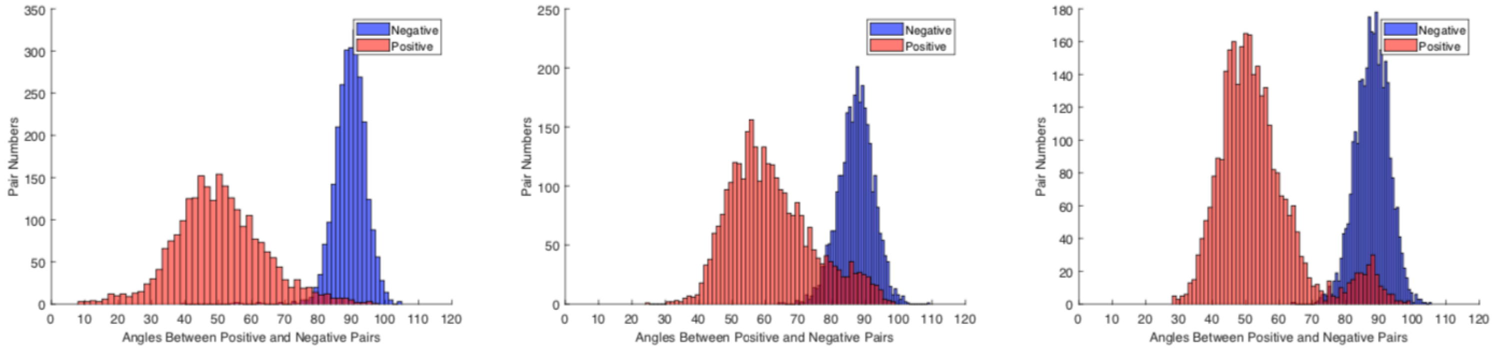
(b) Triplet-Loss

ArcFace has a more distinct margin than Triplet-Loss

Angle distributions on Test Datasets



(a) LFW (99.83%) (b) CFP-FP (98.37%) (c) AgeDB (98.15%)



(d) YTF (98.02%) (e) CPLFW (92.08%) (f) CALFW (95.45%)

Verification Performance (LFW)

Method	LFW	CALFW	CPLFW
HUMAN-Individual	97.27	82.32	81.21
HUMAN-Fusion	99.85	86.50	85.24
Center Loss [38]	98.75	85.48	77.48
SphereFace [18]	99.27	90.30	81.40
VGGFace2 [6]	99.43	90.57	84.00
MS1MV2, R100, ArcFace	**99.82**	**95.45**	**92.08**

Verification Performance (MegaFace)

Methods	Id (%)	Ver (%)
Softmax [18]	54.85	65.92
Contrastive Loss[18, 32]	65.21	78.86
Triplet [18, 29]	64.79	78.32
Center Loss[38]	65.49	80.14
SphereFace [18]	72.729	85.561
CosFace [37]	77.11	89.88
AM-Softmax [35]	72.47	84.44
SphereFace+ [17]	73.03	-
CASIA, R50, ArcFace	77.50	92.34
CASIA, R50, ArcFace, R	91.75	93.69
FaceNet [29]	70.49	86.47
CosFace [37]	82.72	96.65
MS1MV2, R100, ArcFace	81.03	96.98
MS1MV2, R100, CosFace	80.56	96.56
MS1MV2, R100, ArcFace, R	98.35	98.48
MS1MV2, R100, CosFace, R	97.91	97.91

- **Id:** rank-1 face identification accuracy
- **Ver:** face verification TAR at 10^{-6}
- **R:** data refinement on both probe set.

Appendix

Feature Space Analysis.

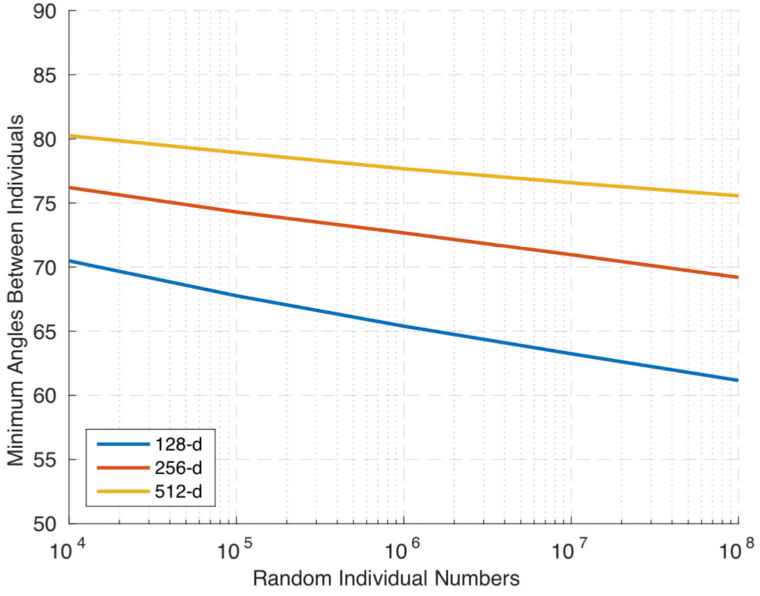
Is the 512-D Enough to hold identities ?

Assume the identity center W_j's follow a realistically spherical uniform distribution,

Expectation of the nearest neighbor separation,

$$\mathbb{E}[\theta(W_j)] \to n^{-\frac{2}{d-1}}\Gamma(1+\frac{1}{d-1})(\frac{\Gamma(\frac{d}{2})}{2\sqrt{\pi}(d-1)\Gamma(\frac{d-1}{2})})^{-\frac{1}{d-1}},$$

$$\theta(W_j) = \min_{1\le i,j\le n, i\ne j} \arccos(W_i, W_j) \forall i, j.$$



The high-dimensional space is so large that $\mathbb{E}[\theta(W_j)]$ decreases slowly when the class number increases exponentially.

Theoretically, Yes.

Thanks.

☺